BIOINFORMATICS

GNARE – A Grid-based Server for the Analysis of User Submitted Genomes

Dinanath Sulakhe [1], Mark D'Souza[2], Mustafa Syed[1], Alexis Rodriguez[1], Yi Zhang[3], Elizabeth M. Glass[1,2], Margaret F. Romine[4], and Natalia Maltsev[1,2*]*

[1]*Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA.* [2]*Computation Institute, University of Chicago, Chicago, IL 60637, USA.* [3]*University of Illinois at Chicago, Chicago, IL 60607, USA.* [4]*Pacific Northwest National Lab, Richland, WA 99352, USA*

ABSTRACT

GNARE (GeNome Analysis Research Environment) is a bioinformatics server that supports both automated and expert-driven (interactive) analysis of user-submitted genomes and metagenomes. These analyses include gene function prediction and development of organism-specific metabolic reconstructions from sequence data. GNARE also provides a framework for comparative and evolutionary analysis as well as annotation of genomes and metabolic networks in the context of phenotypic and taxonomic information. Results of analyses and metabolic models are visualized and extensively annotated with information from public databases. GNARE uses automated workflows and a Grid-based computational backend, to perform high-throughput analysis of genomes. The use of distributed computing in GNARE allows the analysis of an average-sized prokaryotic genome in less than five hours.

Contact: maltsev@mcs.anl.gov, sulakhe@mcs.anl.gov

1 INTRODUCTION

In the past 10 years the amount of data in genomic databases has doubled each year. In order for researchers to take advantage of the vast scientific value of this information for understanding biological systems, the information must be integrated, analyzed, and modeled computationally in a timely fashion. The development of predictive computational models of an organism's functionality is essential for the progress of such fields as medicine, biotechnology, and bioremediation. These models allow for the prediction of protein functions in newly sequenced genomes, as well as the existence of particular metabolic pathways and regulatory networks. Conjectures developed during genome analysis provide invaluable assistance to researchers in experimental planning and help conserve time and resources required for characterizing an organism's biochemical and physiological properties. Essential for fulfilling this task is the development of high-throughput computational environments that integrate (i) large amounts of genomic and experimental data; (ii) comprehensive tools for knowledge discovery and data mining; and (iii) comprehensive user interfaces that provide tools for easy access, navigation, visualization, and annotation of biological information.

Motivation. A significant number of sequencing projects (often for sequencing of a single genome) and initial interpretation of genomes are conducted by universities and small sequencing facilities that do not have sufficient resources for developing highly integrated and scalable bioinformatics systems for the interpretation of newly sequenced genomes. This time and resource consuming task can be afforded only by large bioinformatics centers (Fig. 1). To address the needs of these groups, as well as groups interested in analysis of a particular organism or organisms (e.g., *Shewanella Federation*, biodefense, and MetaGenome projects), we have developed a Web-based public server, GNARE (GeNome Analysis Research Environment). GNARE allows scientific groups and individual users to analyze genomic data using an integrated and automated bioinformatics environment based on advanced computational technologies.

GNARE leverages the following bioinformatics systems and analytical tools being developed by our team:

(1) The PUMA2 system [Maltsev et al, 2006] for high-throughput genetic sequence and evolutionary analyses of genomes and metabolic reconstructions from sequence data. This system contains precomputed analyses of over 1,000 publicly available genomes and automated metabolic reconstructions for over 330 organisms. PUMA2 has been used by the scientific community worldwide.

[*] To whom correspondence should be addressed.

(2) Tools for high-resolution comparative and evolutionary analysis of genomes developed by our group (e.g., Chisel [Chisel, 2006], a workbench for identification of taxonomic and phenotypic variations of enzymes, tools for comparative analysis of metabolic pathways).

(3) GADU [Sulakhe et al, 2005] [Rodriguez et al, 2003] (Genome Analysis and Database Update system),an automated scalable computational pipeline for data acquisition and analysis by a variety of bioinformatics tools. GADU utilizes Grid resources for high-throughput computations (Section 3.4).

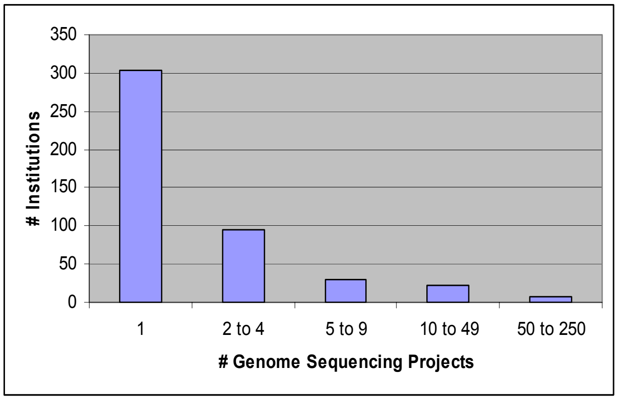

Fig 1: The number of sequencing projects genomes per institution. Note that 66% of institutions have only one sequencing project, while over 87% have four or fewer [Benson et al, 2005].

A number of excellent genome annotation systems are available that have capabilities similar to GNARE. These include ERGO (Integrated Genomics, Inc.) [Overbeek et al, 2003], Pedant-Pro (Biomax Informatics AG) [Biomax], and Phylosopher (Gene Data, Inc.) [Genedata, 2001], but these are commercial applications and are not available freely. Other systems such as Manatee (TIGR) [Manatee, 2005] and SEED [Overbeek et al, 2005], GenDB [Meyer et al, 2003] and BASys [Van Domselaar et al, 2005] are freely available but often require extensive effort to successfully install a system that is fully functional. Installing genome annotation systems locally gives the user complete control of the annotation process and ensures privacy of data, but the user is also responsible for providing the required computational resources. To complete the automated analysis in a timely fashion generally requires a dedicated cluster. GenDB and BASys [Van Domselaar et al, 2005] provide Web applications that permit users to submit genomes for automated annotation utilizing local computational resources.

GNARE has a number of features that distinguish it from the above mentioned systems: (a) GNARE is the only system that utilizes scalable Grid resources as its computational backend, resulting in a considerably shorter turnaround time. GNARE takes under 5 hours to complete the annotation of a 5 Mbp prokaryotic genome. (b) The GNARE system enables users to submit a proteome for analysis, spanning from function prediction to the development of complex metabolic reconstructions. The entire analysis is completed in a simple, transparent, automated fashion. (c) The results are presented to the user in an interactive environment for manual annotation via simple graphic user interface. The user can further explore functional evidence using over thirty external, public, Web-based bioinformatics tools (d) GNARE also supports web-based community curation of genomes.

The following sections describe GNARE's capabilities and architecture in more detail.

2 ANALYSIS OF GENOMES IN GNARE

The major steps of automated analysis of genomes in GNARE and requirements for data integration for their support are presented in Figure 2. As the figure indicates, GNARE currently supports the following analyses: (a) assignment of functions to genes, (b) the development of metabolic reconstructions from sequence data, and (c) comparative analysis by a variety of bioinformatics tools in the framework of phenotypic and taxonomic information. The results of the automated analyses by GNARE may be further refined by the users interactively. GNARE supports both public and private curation of genomes and metabolic models by individual users or groups of experts.

Steps of genome analysis in GNARE

Step 1. Submission of genomes. Users can submit sequence data for analysis via the Web interface or FTP. If their genome of interest has been publicly released and is available in PUMA2, it can be instead selected from the Web-based user interface and used for development of the user's model in GNARE. Currently GNARE accepts translated ORFs in FASTA format for analysis. The next release, anticipated in April 2006, will also accept genomic sequences and DNA contigs. The additional analysis will include prediction of potential ORFs using Critica [Badger et al, 1999], Glimmer [Delcher et al, 1999] or GenMark [Lukashin et al, 1998], according to the user's choice and the visualization of the resulting ORFs in the genomic context.

Step 2. Assignment of function to gene products. GNARE uses a voting algorithm developed by our group to assign potential protein functions. The voting algorithm [manuscript in preparation] utilizes the results of precomputed analysis of sequences by several bioinformatics tools. The results of analyses of the user-submitted data by BLAST [Altschul et al, 1990], Blocks [Henikoff et al, 2000], and Chisel [Chisel, 2006] are computed by GADU (Section 3.4).

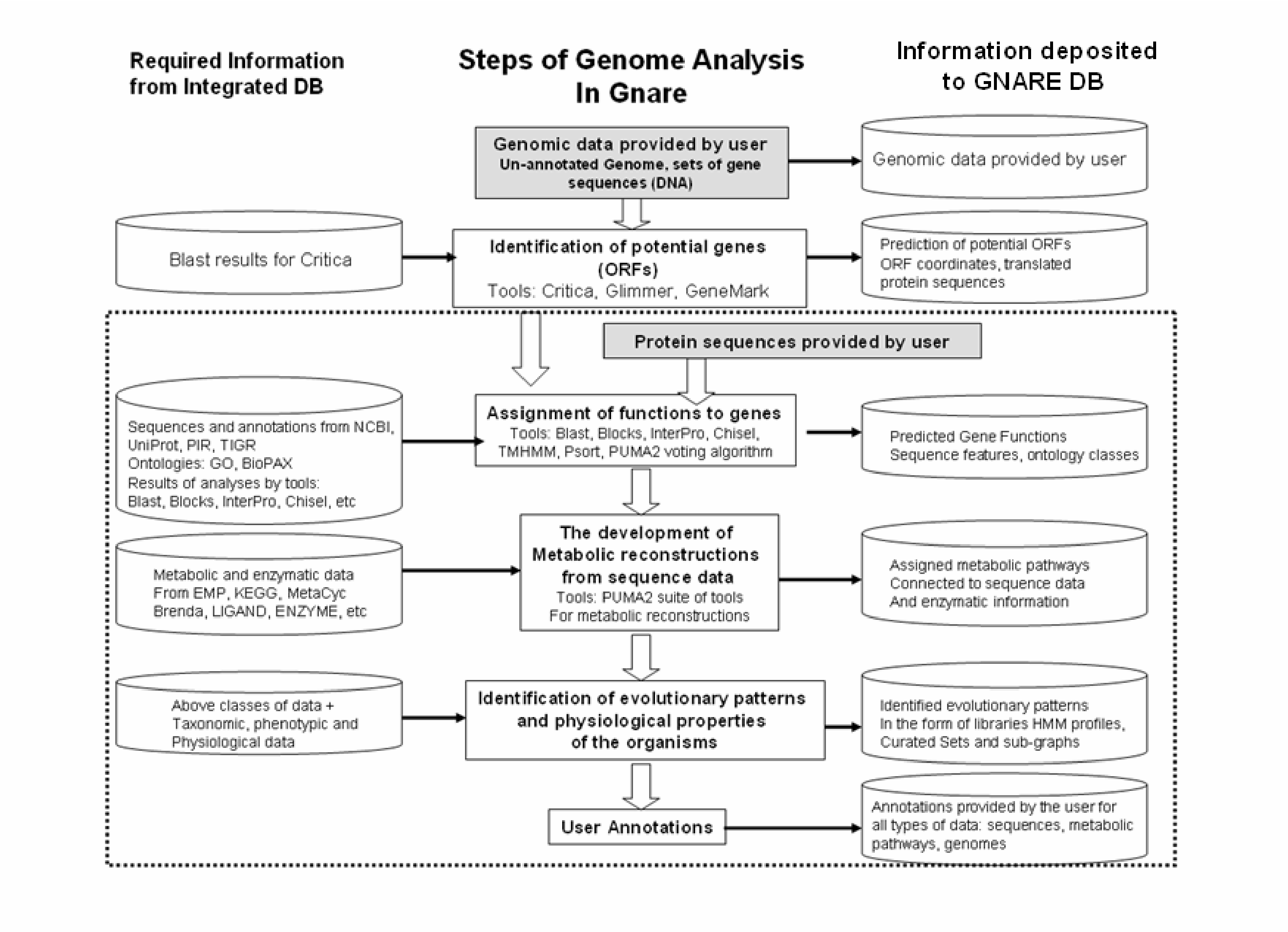

Fig 2: Major steps of automated analysis of genomes in GNARE. Capabilities of the current system is outlined in the dotted box. Entry points for user data are highlighted in grey.

The results of gene function predictions and precomputed results are presented in an interactive interface that supports user curation of every protein sequence annotation. Users can also perform further analysis of the sequence with over 30 bioinformatics tools to enhance their ability to produce accurate gene function predictions. GNARE also facilitates the comparisons of gene function predictions performed by different groups or users. This feature is important for community genome curation. Figure 3 presents a snapshot of GNARE user interface for analysis of protein sequence in *Shewanella oneidensis* MR-1.

Step 3. Metabolic reconstructions from sequence data. Metabolic reconstructions from sequence data (EMP [Selkov et al, 1996], WIT2 [Overbeek et al, 2000], KEGG [Kanehisa et al, 2004], Ecocyc [Keseler et al, 2005] and PUMA2 [Maltsev et al, 2006]) have proved useful for developing organism and process-specific functional models. Identification of gene functions allows reconstruction of metabolic networks that potentially exist in the organism. For example, the presence of all the genes known to be involved in a glycolytic pathway in a proteome suggests the possibility that this pathway may be carried out by the organism under consideration. Such metabolic models represent valuable scientific hypotheses regarding an organism's physiology and facilitate experimental planning. However, the development of metabolic reconstructions requires the integration of sequence data, annotations, and metabolic and taxonomic information in one coherent framework. GNARE leverages the PUMA2 knowledge base for support of the development of automated metabolic reconstructions from the sequence data provided by user. GNARE also supports *interactive* development and annotation of metabolic models. The developed metabolic reconstructions are based on pathway data from the EMP collection of enzymes and metabolic pathways, developed by the EMP Project Inc. [Selkov et al, 1996]. This database contains enzymatic information and metabolic diagrams accumulated from the literature describing over 4,000 metabolic pathways in a structured, indexed, and searchable form.

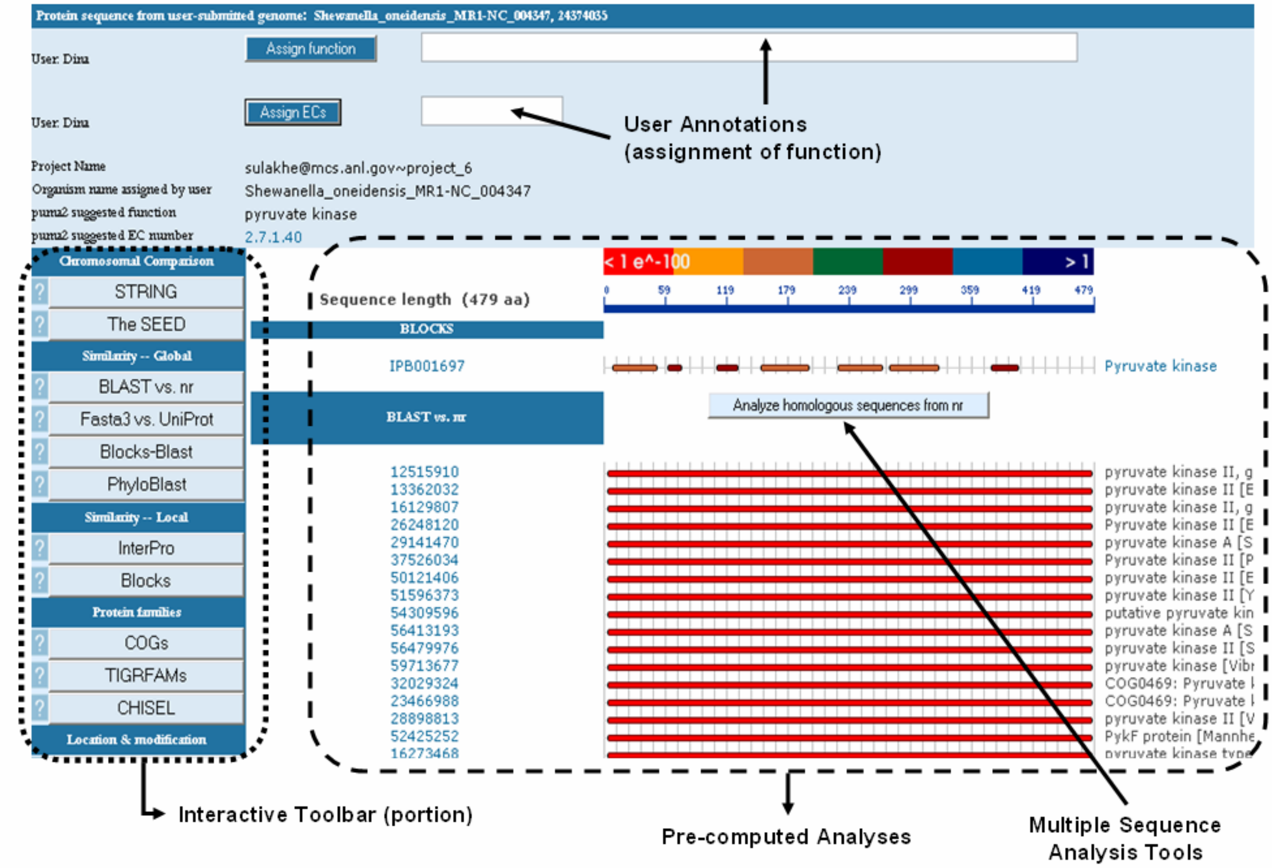

Fig 3: Protein analysis page for pyruvate kinase in *Shewanella oneidensis*. Precomputed analyses are presented for BLAST and Blocks along with interactive sequence analysis tools (only a partial list is shown), multiple sequence analysis tools and the user annotation tool.

In addition to developing automated metabolic reconstructions, GNARE provides tools for identification of “missing” enzymes and for navigation of pathway data in the larger context of biological processes. It supports comparative analysis of metabolic networks that allow identification of variations of the metabolic pathways characteristic to particular organisms or taxonomic groups of organisms. GNARE metabolic reconstructions provide links to navigate through sequence and enzymatic data. Genomic data and metabolic models in GNARE are annotated with the Gene Ontology (GO) [Ashburner et al, 2000] terms and are available in BioPAX [BioPAX, 2006] format.

Step 4: Comparative analysis of genomes in the framework of taxonomic and phenotypic information. Comparative and evolutionary analysis of genomes and metabolic networks in the framework of phenotypic and taxonomic information provides an organism-centric, systems-level view of genomic data. It allows the user to trace the evolutionary history associated with entire biochemical pathways and biological processes, rather than of individual genes, and to reconstruct the evolutionary progression of organisms that possess these pathways. It also enables identification, analysis and characterization of evolutionary patterns associated with particular phenotypes or phylogenetic neighborhoods. GNARE leverages the following technologies developed by our group to support comparative analysis of user-submitted genomes:

(1) Chisel (manuscript in preparation): an integrated computational workbench for identification and characterization of enzymatic sequences. Chisel utilizes information from the PUMA2 integrated knowledge base (Section 3.2) to perform rules-based clustering and classification of annotated enzymatic sequences into functional categories. The resulting clusters are used for the development of a library of Hidden Markov Models (HMM profiles) for particular enzymatic functions. When the data is sufficient, taxonomic and phenotypic variations of enzymes (e.g., viral, gamma-proteobacterial, fungal) are also identified. The HMM profiles are also used for prediction of functions of hypothetical proteins. Currently, the Chisel library contains over 5000 function-specific models for 1242 distinct enzymatic functions. Each model includes an HMM profile, PSSM model, Blocks profiles [Henikoff et al, 2000], as well as ClustalW [Higgins et al, 1998] alignments and a library of degenera-

tive PCR primers and oligonucleotides for the needs of experimentalists**.** The Chisel library of HMM profiles is used in GNARE for identification and characterization of enzymatic sequences in genomes provided by users. Identification of taxonomy-specific variations of enzymes is especially important for interpretation of metagenomic data.

(2) PhyloBlocks: a tool for interactive development of HMM profiles from the sets of homologs selected by the user.

(3) Tools for comparative analysis of metabolic networks.

GNARE supports comparative analysis of metabolic pathways identified by the user in a course of metabolic reconstructions with the pathways from over 300 other organisms. The user is presented with a spreadsheet showing the enzymatic content for a particular pathway in the organisms of their choice. Users are also able to compare pathways in each organism and search for missing enzymes.

3 SYSTEMS ARCHITECTURE

GNARE's architecture (shown in Fig 4) includes the following components:

(1) Web interface for
- user authentication,
- submitting raw genomic data,
- creating workflows for the analysis using various bioinformatics tools,
- monitoring the status of analysis,
- visualizing the results of analysis, and
- user annotations.

(2) Integrated database for user data,
(3) Workflow execution engine,
(4) Interface to the high-throughput Grid computational resources via GADU.

3.1 Interface for User Interactions

All the transactions in GNARE are user specific and can be accessed only by the user through login. An authentication model using login-based access is used for validating all the users and their transactions on the GNARE server. The user-submitted input data, the analysis performed by the user, and the results are stored in a user-specific space.

Users can upload the raw protein sequence data in the form of a FASTA file through the Web interface, or they can provide an FTP path to the input data. The interface provides a Web form where the user can select the bioinformatics tools for the analysis of uploaded data. Currently GNARE supports Grid-based analysis using BLAST, Blocks, and Chisel. This set of tools will be expanded in the near future. The user can monitor the progress of the analysis as these tools are being executed by the Workflow Executor (Section 3.3). After Grid-based analysis of the data by the bioinformatics tools is completed, the data is further analyzed by the gene function predictions algorithm. The results of gene function predictions are used for the development of an automated metabolic reconstruction. The results of all analyses are visualized and presented to the user through the Web interface. GNARE uses features of the PUMA2 system for representation of the analysis results. Various templates are used from the framework of PUMA2, such as the protein page representing BLAST and Blocks results, metabolic reconstruction, and comparison of metabolic pathways.

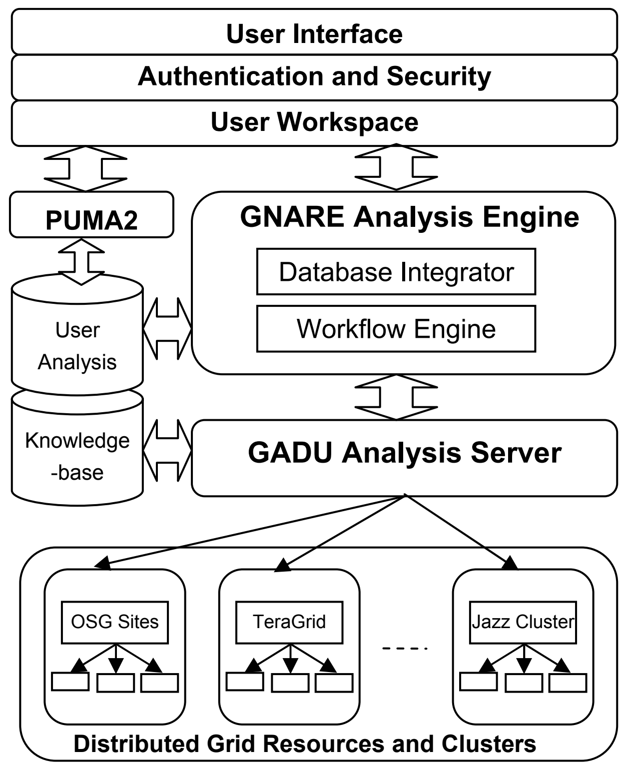

Fig 4. Components of GNARE architecture and their interactions.

3.2 Integrated User Database

Efficient comparative analysis in GNARE is supported by integrated knowledge base consisting of two parts:

(1) A relational integrated PUMA2 database containing vast amounts of genomic, metabolic, enzymatic, taxonomic and phenotypic information obtained from over 20 public databases and the precomputed results of analyses by various bioinformatics analysis tools (e.g. BLAST, Blocks and InterProScan [Mulder et al, 2005]) for over 1000 genomes.

(2) A relational database for user-submitted sequence data, results of analyses by the bioinformatics tools,

metabolic reconstruction, functional predictions, and user annotations.

3.3 GNARE Analysis Engine

The GNARE analysis engine executes the steps described in Section 2 for the interpretation of user-submitted genomes. The initial data submitted, as well as the intermediate data generated during the analysis is stored in the user's workspace. The analysis engine has two modules: a database integrator and a workflow engine for user-submitted data. The *database integrator* stores the sequence data submitted by the user and the data generated by the tools at various levels of analysis into the relational user database. The *workflow engine* executes the various analysis tools in the specified order, taking care of the dependencies.

3.4 High-Throughput Grid Backend

Analyses of large volumes of data by a variety of bioinformatics tools require substantial computational resources. Also required are automated workflows to ensure the reliability and stability of the execution of analytical steps. GADU is an automated, scalable, high-throughput computational workflow engine that enables the Grid execution of bioinformatics tools. It acts as a gateway to the Grid, handling all computational analyses for the GNARE system.

GADU has access to thousands of CPUs from various large-scale Grid resources such as the Open Science Grid (OSG) [OpenScienceGrid, 2006] and TeraGrid (TG) [Teragrid, 2006]. GADU's flexible architecture makes it simple to use Grid resources of different architectures [Sulakhe et al, 2005], such as 32-bit processors of OSG and 64-bit processors of TG and can easily add new resources to its pool. GADU can run its jobs on stand alone clusters as well, such as the Jazz cluster [LCRC, 2006] at Argonne National Laboratory. GADU uses the Grid resources on an opportunistic basis with no reservations for resources. Whereas using shared local clusters requires reservations for faster results.

4 RESULTS

GNARE has been used for analysis of 11 Shewanella strains for the needs of *Shewanella Federation*, the Hanford site microbial community MetaGenome, and genomes of pathogenic organisms for the NIH GLRCE for Biodefense and Emerging Infectious Disease Research Consortium.

The *Shewanella Federation* (DOE OBER) aims to characterize and model the biology of *Shewanella oneidensis* MR-1 as well as other members of this Genus. This facultative anaerobe is capable of using a diverse array of electron acceptors (nitrite, nitrate, trimethylamine oxide, fumarate, dimethyl sulfoxide, thiosulfate, iron, elemental sulfur, manganese, and uranium) to support growth in the absence of oxygen. Comparitive genome analyses of *Shewanella* species isolated from different sites around the world presents a unique opportunity to explore the link between their functional make-ups with their ability to thrive in varied ecological niches.

Genomes of 11 strains and an MR-1 plasmid of Shewanella (total 43,839 protein sequences) were submitted to GNARE via the Web interface. This data was analyzed by the following bioinformatics tools: BLAST against the NCBI non-redundant database and Shewanella genomes, Blocks, Chisel, and the function prediction voting algorithm. Automated metabolic reconstructions for all 11 genomes were also developed. The analysis was performed on 1317 CPUs of distributed computational resources from OSG and TG.

GNARE offered the following advantages to the Shewanella research community:

a. *Fast analysis of genomes*. Analysis of all 11 strains of Shewanella from submission to display of the visualized results and metabolic reconstructions via the interactive Web interface took 64 hours.

b. *Efficient comparative analysis*. Simultaneous analysis of several strains of *Shewanella* (including ones not yet available in the public databases) and comparative analysis with genomes of over 1,000 organisms from the integrated knowledge-base provides a powerful environment for refinement of the ORF boundaries, improvement of gene function prediction, identification of candidate essential genes, and evolutionary analysis of *Shewanella* sequences in the framework of taxonomic and metabolic information.

c. *Metabolic reconstructions.* Integration of metabolic and taxonomic data in GNARE's framework provides a unique systems-level perspective of an organism. The development of metabolic reconstructions is especially important for the researchers of *Shewanella Federation* because of their primary interest in the physiology and metabolism of these organisms. Comparative analysis of metabolic pathways allows the detection of metabolic diversity between various strains of *Shewanella.* Identification of candidates for "missing" enzymes using high-resolution analysis of sequences in Chisel and PhyloBlocks aids in developing of conjectures regarding gene functions and planning experiments.

d. *Support of interactive analysis and community curation.* The *Shewanella* Federation is a multi-investigator, cross-institutional consortium. A collaborative environment that allows sharing and comparing of results of analyses by various researchers and scientific groups is essential for the success of the project. Currently GNARE offers comparisons of the gene functions predictions for *S. oneidensis* MR-1 performed by NCBI, TIGR, PUMA2, the gene function prediction algorithm and Chisel, as well as by manual curation by the members of *Shewanella* consortium. The researchers can also record their notes and suggestions relevant to analyses and share them with the others.

Similar analysis was performed for the Hanford site MetaGenome. An essential part of this analysis was identification of taxonomic variations of enzymes using Chisel. Such analysis allows developing or refining suggestions regarding the taxonomy of the organisms found in the metagenome and attributing particular enzymatic steps to specific taxonomic groups. The results of this analysis are available at the GNARE site.

Functional and metabolic models were created to support research projects for the NIH GLRCE for Biodefense and Emerging Infectious Disease Research consortium. *Bacillus anthracis* strains were analyzed using GNARE for the identification and characterization of essential genes. These essential gene candidates were then used in therapeutic inhibition studies.

5 CONCLUSIONS AND FUTURE PLANS

The availability of new genomic DNA sequences and tools for identifying and predicting functions encoded therein are growing rapidly. To effectively utilize the value of this information for developing a systems level understanding of individual organisms and microbial communities, high-throughput computing and data integration technologies are required that can accommodate the growth of the data and increasing requirements for its analysis . GNARE provides a collaborative environment for comprehensive analysis of genomic information based on advanced computational technologies such as (distributed computing, automated controlled workflows, and data provenance). It offers unique invaluable bioinformatics services, such as the interactive development of metabolic reconstructions, comparative analysis of sequences and metabolic networks, all in the framework of taxonomic and phenotypic information. In the future, GNARE will also accept genomic sequences and DNA contigs as input for the analysis. We plan to allow annotation of genomes with additional classes of information submitted by the users (e.g. gene expression data, biochemical and enzyme kinetic data), increase the number of Grid-supported bioinformatics tools for high-throughput analysis of data, implement Web-services to access additional tools, and a Web portal to provide efficient collaborative environment.

AVAILABILITY

GNARE is a public system available at http://compbio.mcs.anl.gov/gnare. Users can access GNARE with a guest login to browse the precomputed results. In order to perform high-throughput analysis user registration is required. Contact the administrator at gnare@mcs.anl.gov for further details.

ACKNOWLEDGMENTS

N. Maltsev, E. M. Glass, D. Sulakhe, and M. Syed acknowledge membership in and support from the Region V "Great Lakes" Regional Center of Excellence in Biodefense and Emerging Infectious Diseases Consortium (GLRCE, National Institute of Allergy and Infectious Diseases Award 1-U54-AI-057153). M. D'Souza acknowledges membership in and support from the NMPDR Bioinformatics Resource Center NIH/NIAID (Award NNSN 266200400042C). This work was supported in part by the Office of Biological and Environmental Research, US Department of Energy, under Contract W-31-109-Eng-38.

The GNARE development team is grateful for the help, advice and invaluable contributions from the ANL Globus group, especially Mike Wilde, Veronika Nefedova and Ian Foster; former members of the team, Luke Ulrich, Jason Ting and Tanuja Bompada; and Robert Petryszak from the European Bioinformatics Institute for making ClustR available for our use.

REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet. 25*, 1 (2000), 25--29.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. Basic local alignment search tool. *J Mol. Biol. 215*, (1990), 403--410.

Badger, J.H., Olsen, G.J. CRITICA: Coding Region Identification Tool Invoking Comparative Analysis. *Mol. Biol. Evol. 16*, 4 (1999), 512–524.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L. Genbank. *Nucleic Acids Res., 33* (2005), 34-38

BioPAX Home page – Biological Pathways Exchange, 2006 <http://www.biopax.org>

Chisel, high-resolution evolutionary analysis of enzymatic sequences Argonne National Laboratory, 2006. <http://compbio.mcs.anl.gov/CHISEL>

Delcher, A.L., Harmon, D., Kasif, S., White, O., Salzberg, S.L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res. 27*, 23 (1999), 4636-4641.

Teragrid, Distributed Terascale Facility, <http://www.teragrid.org>

Henikoff, J.G., Greene, E.A., Pietrokovski, S., and Henikoff, S. Increased coverage of protein families with the blocks database servers, *Nucl. Acids Res. 28*, (2000), 228--230.

Higgins, D., Thompson, J., Gibson, T., Thompson, J.D., Higgins, D.G., Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res. 22*, (1998), 4673-4680.

LCRC, the Argonne National Laboratory Computing Project, 2006 <http://www.lcrc.anl.gov/jazz/index.php>

Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham,J., Paley, S., Paulsen, I.T., Peralta-Gil, M., Karp, K. EcoCyc: A comprehensive database resource for *Escherichia coli*. *Nucleic Acids Research*, 33, (2005), D334-7.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res. 1*, 32 (2004), D277--280.

Lukashin, A.V., Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res. 26*, 4 (1998) 1107–1115.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Pagni, M., Ponting, C.P., Quevillon, E., Selengut, J., Sigrist, C.J., Silventoinen, V., Studholme, D.J., Vaughan, R. and Wu, C.H. InterPro, progress and status in 2005. *Nucleic Acids Res. 33*, (2005), D201--205.

Manatee – The Institute for Genomic Research (TIGR), 2005 <http://manatee.sourceforge.net/ >

Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., Puhler, A. GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res. 31*, 8 (2003), 2187--2195.

Maltsev, N., Glass, E., Sulakhe, D., Rodriguez, A., Syed, M.H., Bompada, T., Zhang, Y., and D'Souza, M. PUMA2 – Grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res. 34*, (2006), D369-372.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., Vonstein, V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 33, 17 (2005), 5691-5702.

Overbeek, R., Larsen, N., Pusch, G.D., D'Souza, M., Selkov Jr, E., Kyrpides, N., Fonstein, M., Maltsev, N., Selkov, E. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res., 28, 1,* (2000), 123–125.

Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov, E. Jr., Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I., Bhattacharyya, A., Burd, H., Gardner, W., Hanke, P., Kapatral, V., Mikhailova, N., Vasieva, O., Osterman, A., Vonstein, V., Fonstein, M., Ivanova, N., Kyrpides, N. The ERGO genome analysis and discovery system. *Nucleic Acids Res. 31*, 1 (2003) 164--171.

OpenScienceGrid.org, 2006 <http://www.opensciencegrid.org>

Biomax Informatics AG: The Pedant-Pro Sequence Analysis Suite <http://www.biomax.com/products/pedantpro/pedantpro.htm>

Genedata Phylosopher 3.5, 2001, <http://www.genedata.com/press/pressreleases/2001/phylosopher/index_eng.html>

Rodriguez, A., Sulakhe, D., Marland, E., Nefedova, V., Yu, G.X., Maltsev, N. GADU - Genome Analysis and Database Update Pipeline. *ANL/MCS-P1029-0203*, (2003).

Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L., Selkov, E. Jr and Yunus, I. The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. *Nucleic Acids Res. 24*, 1 (1996), 26--28.

Sulakhe D, Rodriguez A, D'Souza M, Wilde M, Nefedova V, Foster I, Maltsev N. Gnare: automated system for high-throughput genome analysis with grid computational backend. *J Clin Monit Comput. 19*, 4-5 (2005), 361-369.

Van Domselaar, G.H., Stothard, P., Shrivastava, S., Cruz, J.A., Guo, A., Dong, X., Lu, P., Szafron, D., Greiner, R., Wishart, D.S. BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res. 1*, 33 (2005), W455-459.